

# AO: An Open Annotation Ontology for Science on the Web

Paolo Ciccarese\*, Marco Ocana\*\*, Sudeshna Das\*, and Tim Clark\*‡

\*Harvard Medical School and Massachusetts General Hospital, Boston MA; \*\*Balboa Systems, Newton MA

## ABSTRACT

We present the Annotation Ontology (AO), an open ontology in OWL for annotating scientific documents on the web. AO supports both human and algorithmic content annotation. It enables “stand-off” (separate) metadata anchored to specific positions in a web document by any one of several methods. In AO, the document may be annotated but is not required to be under update control of the annotator. AO contains a provenance model to support versioning, and a set model for specifying groups and containers of annotation.

## 1 INTRODUCTION

Much current work in biomedical ontologies now focuses on detailed formal classification of objects, functions and processes, using description logics.[1-3] This approach creates a set of fixed categories for searching and navigating ontology-annotated content on the web, whether in the standard journal publications or in web “collaboratories” [4].

However, we currently lack a robust common set of methods for linking text in new scientific publications to ontological elements, with full annotation provenance. Given such a facility, formal ontologies can serve as schemas for extremely rich stores of metadata on web documents, linking new scientific content across scientific specializations and collaborations. One fundamental requirement for such methods, if they are to become widely used, would be a formal specification of its metadata. Seminal lines of research in distributed link services [5] and in conceptual open hypermedia[6] have explored this area, without yet to our knowledge publishing an annotation metadata specification meeting requirements for the semantic web.

Not only subject area ontologies, but also a straightforward annotation ontology and a framework for generating annotations with algorithmic assistance, are all fundamentally required for successfully propagating the collaborations model and reaping its full benefits.

We have developed an annotation ontology specifically designed to support content linking in collaborations. Content in collaborations has the great advantage of providing a strong focus to the collected, evolving discourse. Specialists accessing material in such a focused web community – such as PD Online [7] (<http://pdonlineresearch.org>), or Al-

zforum [8], (<http://www.alzforum.org>) – will not need to wade through extraneous material on cardiology, drug addiction, hematology, and so forth. Essentially what these communities do is dramatically improve the signal-to-noise ratio for specialists, making the information explosion in science nearly tractable within a given specialty.

Annotation – either marking up contributions with comments, or more importantly, with relevant concepts and entities from biomedical ontologies – provides a technological boost to “strategic reading” for members of such communities [9,10] and selectively breach established specialist focus boundaries and semantic barriers where required [11].

Existing ontologies and vocabularies which can serve as a basis for such annotation are particularly abundant in the biomedical field and are often expressed in OWL/RDF [12] or in SKOS [13], with OWL/RDF now apparently the most favored option. Subjects for ontological structuring include biological processes, molecular functions, anatomical and cellular structures, tissue and cell types, chemical compounds, and biological entities such as genes and proteins.

We take proteins as a typical example. There are a number of database resources that catalog and identify proteins. UniProt is certainly the most popular but, at the moment, is not available in OWL format (i.e. as a description logic). The PRO Ontology [14] is a project which represents a growing proportion of the content of UniProt and other protein databases as declarations in OWL, and is interoperable with other OBO Foundry ontologies - such as the Sequence Ontology [15] and the Gene Ontology [16] - that provide representations of protein qualities. This interoperability facilitates cross-species comparisons, pathway analysis, disease modeling, and the generation of new hypotheses through data integration and machine reasoning.

Our annotation ontology was motivated by these requirements and use cases. It has also been influenced by an analysis of strengths and weaknesses of earlier work by Swick et al. in the Annotea Project [17], discussed below.

Annotea was developed as a Web-based shared annotation system based on a general-purpose open RDF infrastructure, where annotations are modeled as a class of metadata. Annotations, identified by a URI, are viewed as statements or remarks made by an author about a Web document. In the Annotea model (see figure 1), the context defines where

‡ To whom correspondence should be addressed: [tim\\_clark@harvard.edu](mailto:tim_clark@harvard.edu)

PAV (Provenance Authoring and Versioning) module of the SWAN ontology [23,24]. Resources targeted by an annotation are usually documents (foaf:Document). In online scientific communities these will be HTML pages, documents and images but many other resources can potentially be annotated, including sections of documents, images and other online resources.



This variety of possible targets is an important requirement motivating introduction of the class Selector. A Selector identifies a portion of a resource, and may work differently for different types of documents and content types. It is also possible to provide different selector models for the same resource type. For instance, for selecting a chunk of text in a XHTML document we can use mechanisms based on XPointer, an offset as in DLS, or other more robust mechanisms. In fact, for immutable content selectors of the type XPointer or offset and length might be easier to deal with. In general, though, it is well known that not all the HTML pages are immutable and some sections of these pages may vary along time - news and advertisements are often embedded in the document - requiring more reliable and customized context definition mechanisms.

[illegible]

### Figure 2 - The Annotation Ontology model

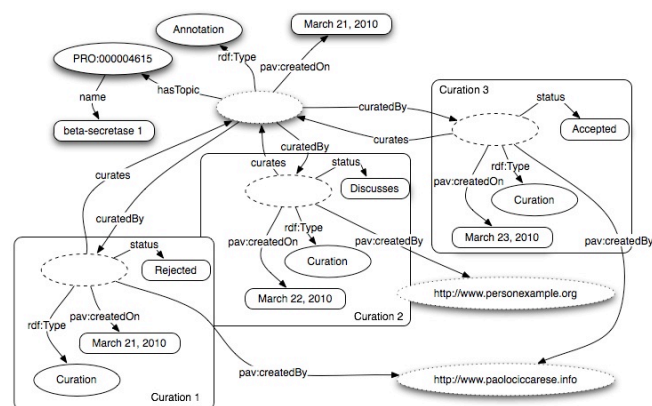
Selector can be subclassed according to the needs to define selectors for images, records in databases and text. In Figure 2, TextSelector – a subclass of Selector - allows us to identify a portion of the document text and link it with a term from a formally defined vocabulary. As in Annotea, it is possible to additionally create sub-types of annotations by sub-classing the Annotation class. Through this mechanism it is possible, for instance, to introduce 'comment' where the purpose is not to attach a term but to attach an explanatory text to a portion of a document. The list of possible subtypes can be virtually unlimited. In distinction to Annotea where users can define types of annotations on the fly, our approach is more conservative, using predefined discourse relationships from the SWAN ontology.

2

### 3 CURATION

Curation is a crucial aspect of scientific publication and therefore an important aspect for our annotation ontology. We enable a curation process for annotation generated by both humans and text mining services. Figure 2 demonstrates an annotation generated by a text mining service, which is then evaluated and accepted by a human curator. In general, every annotation can undergo a multi-step curation process that can involve one or more users.

In figure 3 we show a typical example of semi-automatic annotation workflow that can be summarized as follows: annotation is created by a text mining service, and a user expresses a judgment on the validity of the automatic annotation, for instance "rejected" (curation #1). Later on a second curator might want to discuss the reason(s) for rejection



(curation #2 with status: discussed). And finally a decision

**Figure 3 - Multiple curation steps applied to figure 2 annotation. Some details have been removed for compactness of presentation.**

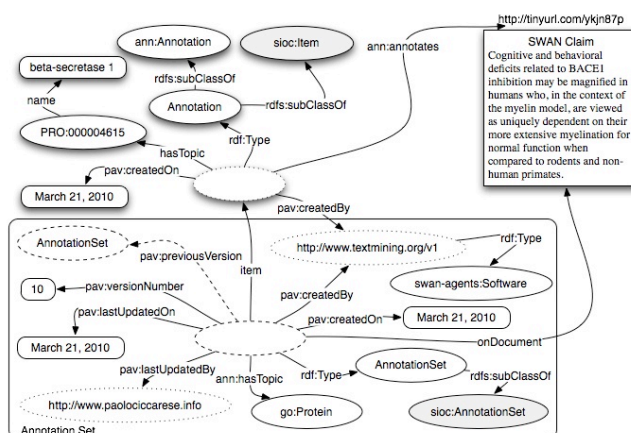
is taken and the annotation is either rejected or accepted (curation #3 with status: accepted). We are assuming the curation process to be a linear story where the timeline can be determined through the curation dates. Alternatively, it is possible to compile explicitly an ordered list of curation item using for instance the SWAN Ontology module for collections.

### 4 ANNOTATION SETS

Often it is particularly useful to be able to group annotations by topic or through other criteria such as all the officially curated annotations. Always thinking of text mining services, there might be a text mining service able to provide multiple types of annotations: genes, proteins and so on.

The provenance of such annotation is always the same but the topics of the annotation are different. We might want to group all the annotations with topic gene. For doing so we

introduced the concept of Annotation Set. The Annotation Set is a container of annotations.



**Figure 4 - Annotation Set**

In figure 4 besides the annotation set and its provenance and versioning it is also possible to detect the mappings of the annotation ontology to the SIOC ontology. As mentioned above, SWAN and SIOC have been the objects of an alignment process in the context of the Scientific Discourse Task Force, one of the sub groups of the W3C Health Care and Life Sciences Working Group. As creators of the SWAN ontology we confirm our commitment in keeping the two efforts aligned.

The AO model of scientific document annotation has been tested in prototype and early-development versions of the SWAN Annotation Framework for both manual and machine-generated markup of web content.

### 5 CONCLUSIONS

This model of web document annotation permits users including journal or web community editorial staff, individual scientists, and computational web agents to construct and persist scientific document annotation as RDF, linking text strings within the document to term URIs in scientific – particularly biomedical – ontologies. It supports multiple methods for linking terms to specific locations in documents, and depending upon the method used, can be stored entirely independently of the target document, which itself can remain unchanged.

The AO ontology provides essential abilities for scientific web communities and publishers, including support for: (1) building position-aligned term enrichment into documents regardless of whether or not one controls the original text; (2) linking content across web communities and communi-

ties of scientific users, with shared metadata; (3) constructing searchable semantic metadata stores linked to documents in a standard way; (4) curation, with provenance, authoring and versioning of all annotations; and (5) human, algorithmic, and human-reviewed algorithmic annotation. AO is available as an OWL file at <http://purl.org/swan/annotation>.

## 6 REFERENCES

- [1] Smith, B. (2004) The Logic of Biological Classification and the Foundations of Biomedical Ontology. In *Invited Papers from the 10th International Conference in Logic Methodology and Philosophy of Science*, Oviedo, Spain, 2003 (Westerståhl, D., ed.) Elsevier-North-Holland, Amsterdam.
- [2] Smith, B. et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25, 1251-5.
- [3] Alexander, C.Y. (2006). Methods in biomedical ontology. *J. of Biomedical Informatics* 39, 252-266.
- [4] Bos, N., Zimmerman, A., Olson, J., Yew, J., Yerkie, J., Dahl, E. and Olson, G. (2007). From shared databases to communities of practice: A taxonomy of collaboratories. *Journal of Computer-Mediated Communication* 12, article 16.
- [5] Carr, L., DeRoure, D., Hall, W. and Hill, G. (1995) The Distributed Link Service: A Tool for Publishers, Authors and Readers. In *Fourth International World Wide Web Conference*. World Wide Web Consortium (W3C), Boston, Massachusetts, USA.
- [6] Bechhofer, S., Yesilada, Y., Stevens, R., Jupp, S. and Horan, B. (2008). Using ontologies and vocabularies for dynamic linking. *IEEE INTERNET COMPUTING* 12, 32-39.
- [7] Das S, Rogan M, Kawadler H, Corlosquet S, S, B. and T, C. (2010) PD Online: a case study in scientific collaboration on the Web. In *Workshop on the Future of the Web for Collaborative Science*, 19th International World Wide Web Conference, Raleigh, NC, USA.
- [8] Kinoshita, J. and Clark, T. (2007). *Alzforum*. *Methods Mol Biol* 401, 365-81.
- [9] Renear, A.H. and Palmer, C.L. (2009). Strategic Reading, Ontologies, and the Future of Scientific Publishing. *Science* 325, 828 - 832.
- [10] Shotton, D. (2009). Adventures in Semantic Annotation. *PLoS Computational Biology* 5, e1000361.
- [11] Das, S., Girard, L., Green, T., Weitzman, L., Lewis-Bowen, A. and Clark, T. (2009). Building biomedical web communities using a semantically aware content management system. *Brief Bioinform* 10, 129-38.
- [12] McGuinness, D. and van Harmelen, F. (2004) OWL Web Ontology Language. W3C Recommendation.
- [13] Miles, A. and Bechhofer, S. (2009) SKOS Simple Knowledge Organization System Reference. W3C Recommendation.
- [14] Natale, D. et al. (2007). Framework for a Protein Ontology. *BMC Bioinformatics* 8, S1.
- [15] Eilbeck, K., Lewis, S., Mungall, C., Yandell, M., L, S., R, D. and M, A. (2005). The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology* 6, R44.
- [16] Ashburner, M. et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25-9.
- [17] Kahan, J., Koivunen, M.-R., Prud'Hommeaux, E. and Swick, R.R. (2001) Annotea: An Open RDF Infrastructure for Shared Web Annotations. In *WWW10 International Conference*, Hong Kong.
- [18] Becket, D. (2004) RDF/XML Syntax Specification (Revised). World Wide Web Consortium
- [19] Grosso, P., Maler, E., Marsh, J. and Walsh, N. (2003) Xpointer Framework. W3C Recommendation. World Wide Web Consortium.
- [20] DeRose, S., Maler, E. and Orchard, D. (2001) XML Linking Language (XLink) Version 1.0. In W3C Recommendation.
- [21] Fielding, R., Getty, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. and Berners-Lee, T. Hypertext Transfer Protocol -- HTTP/1.1. Internet Engineering Task Force, RFC 2616.
- [22] Brickley, D. and Miller, L. FOAF Vocabulary Specification 0.97. Namespace Document 1 January 2010 - 3D Edition.
- [23] Ciccarese, P. (2009) Semantic Web Applications in Neuromedicine (SWAN) Ontology, W3C Interest Group Note 20 October 2009. World Wide Web Consortium.
- [24] Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A. and Clark, T. (2008). The SWAN biomedical discourse ontology. *J Biomed Inform* 41, 739-51.